Beyond Top Probability: Shannon Entropy as a Confidence Signal in Football Forecasting

Quentin Barbedienne
Applied sports analytics and probabilistic modeling
Historical case study based on Foresportia forecasts

May 2026

Abstract

Football prediction systems often expose the most likely outcome and its probability, but this top probability alone does not characterize the remaining uncertainty in a multi class forecast. This note studies Shannon entropy as a complementary confidence signal for 1X2 football forecasts. Given a calibrated distribution $p = (p_{home}, p_{draw}, p_{away})$, entropy measures how dispersed the full forecast is, while $p_{\max}$ only measures its highest point. Using a historical case-study dataset of 14,650 finished football forecasts, we compare a naive high probability filter, $p_{\max} \geq 0.60$, with a low entropy filter, $H(p) \leq 1.15$ bits. The $p_{\max}$ filter selects 2,157 matches with 1,709 correct top pick outcomes (79.2% observed success), while the low entropy filter selects 759 matches with 671 correct outcomes (88.4%). On the latest 100 matches in each segment, the $p_{\max}$ filter records 62/100, while the low entropy subset records 88/100. The interpretation is deliberately selective rather than universal: entropy does not remove football uncertainty, but it can reveal when a forecast is concentrated enough to be communicated more responsibly.

Keywords: sports forecasting; football prediction; Shannon entropy; uncertainty; calibration; multi class classification; probabilistic forecasting; draw traps.

Disclosure. Quentin Barbedienne is the creator of Foresportia, the system from which the aggregate case-study metrics are derived. This note is an applied sports analytics study about probabilistic forecast readability and uncertainty communication. It does not provide betting advice, financial advice, or individualized prediction recommendations.

1 Introduction

Sport is compelling because uncertainty never fully disappears. A model can identify a strong favorite, and still leave enough probability mass on the draw to make the forecast fragile. A football team can have a 70% win probability and still not be a risk-free prediction. This is not a contradiction. A 1X2 forecast is not a single number; it is a probability distribution over three outcomes:

$$p = (p_{home}, p_{draw}, p_{away}).$$

A forecast such as $p = (0.70, 0.25, 0.05)$ clearly identifies a favorite, but it also assigns one quarter of the probability mass to the draw. In football, where scoring is low and draws are structurally common, that residual mass matters.

Many forecasting interfaces emphasize the top probability, or $p_{\max}$. This is useful, but incomplete. $p_{\max}$ tells us which outcome the model prefers. It does not tell us how the remaining probability

mass is distributed. A favorite at 70% with a 25% draw probability is not equivalent to a favorite at 70% with two alternatives at 15% each.

This note examines a simple idea: use Shannon entropy to measure how concentrated the full 1X2 distribution is. The goal is not to suppress uncertainty, nor to convert probabilistic forecasts into certainty. The goal is to communicate model risk more honestly: when a forecast looks confident, is the whole distribution also concentrated?

The data source is a historical export from Foresportia, a football forecasting system developed by the author. The methodological point is more general: in multi class sports forecasting, the top class probability should not be the only confidence indicator.

2 Background: top probability, margin, and entropy

For a discrete probability distribution $p = (p_1, \ldots, p_K)$, Shannon entropy is defined as [1]:

$$H(p) = -\sum_{i=1}^{K} p_i \log_2(p_i). \tag{1}$$

For a 1X2 football forecast, $K = 3$, so:

$$H(p) = -p_{home} \log_2(p_{home}) - p_{draw} \log_2(p_{draw}) - p_{away} \log_2(p_{away}). \tag{2}$$

The maximum entropy for a three outcome forecast is:

$$H_{max} = \log_2(3) \approx 1.585 \text{ bits}, \tag{3}$$

which occurs when the three outcomes are nearly balanced, for example 33%/33%/34%.

$p_{\max}$, probability margin, and entropy answer related but distinct questions. $p_{\max}$ asks how high the leading probability is. The margin asks how far the leading probability is from the runner-up. Entropy asks how dispersed the entire distribution is. In a multi class problem, the last question is important because two forecasts can share the same $p_{\max}$ while encoding different residual uncertainty.

Figure 1 gives the core intuition. Both forecasts have $p_{\max} = 70\%$, but Forecast A leaves most of its residual mass on the draw, while Forecast B spreads it across draw and away. The leading probability alone does not distinguish these cases.

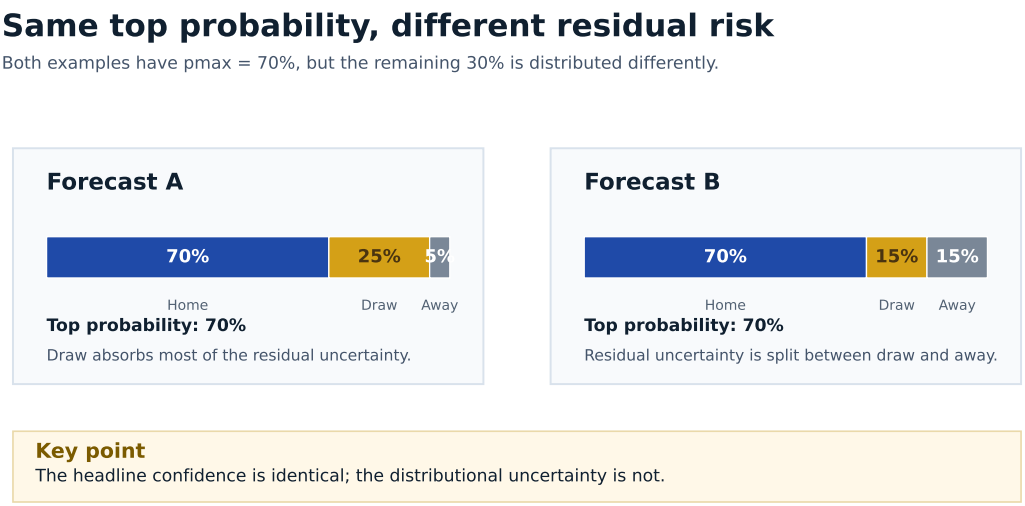

Figure 1: Two forecasts can share the same top probability while carrying different residual risk. In football, the draw probability can dominate the remaining uncertainty behind a favorite.

Figure 2 places this intuition on an entropy scale. A distribution such as 70/25/5 is already concentrated for football, but it is not risk-free: the draw still carries meaningful probability mass. This distinction is central to the article. Entropy is a concentration measure, not a guarantee.

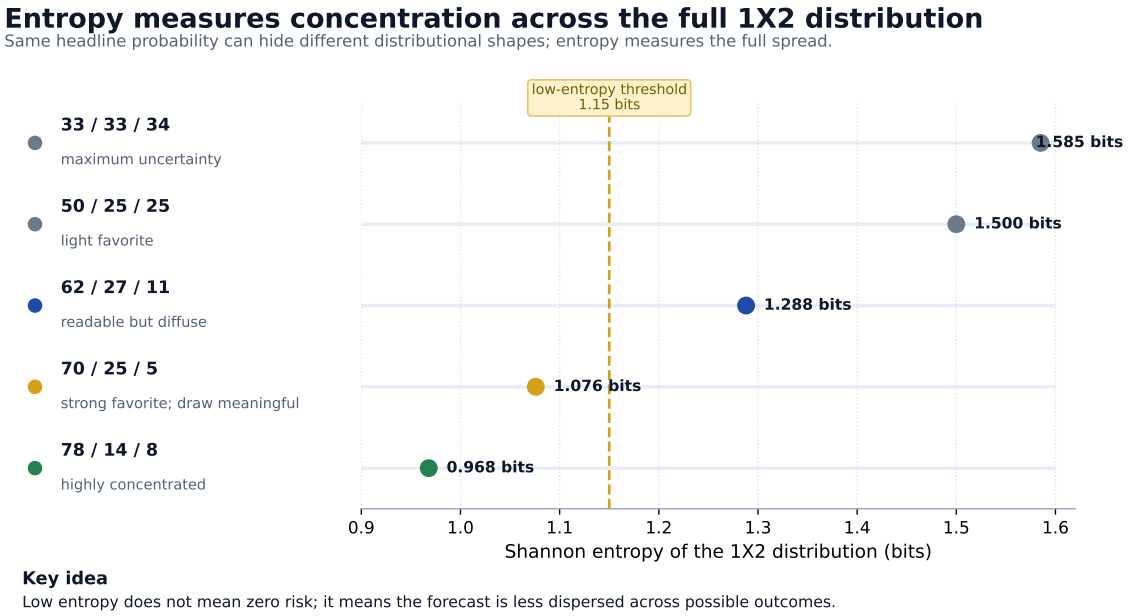

Figure 2: Entropy as a concentration scale for 1X2 forecasts. Lower entropy means a less dispersed forecast, not a risk-free prediction.

3 Why football is a useful test case

Football is a low-scoring sport. A single event can dominate the final result: a red card, a penalty, an early goal, a late equalizer, tactical conservatism, or a favorite failing to convert territorial dominance. As a result, even strong favorites often retain non-trivial residual risk.

This makes football a useful test case for entropy-based confidence signals. In a calibrated football model, entropy below 1 bit is rare. A threshold that may not look extreme in a purely theoretical multi class example can be selective in real football forecasting.

The historical dataset used in this note contains 14,650 completed football forecasts from 2023-09-19 to 2026-05-11. Across the full dataset, the top pick outcome was correct in 7,896 matches, or 53.9%. This global value is not the target metric; it simply describes the full population before any confidence filtering.

The entropy distribution itself explains why the selected threshold is selective. The observed median is 1.536 bits, close to the theoretical maximum of 1.585 bits. Only 5.2% of matches fall at or below 1.15 bits, and only 2.4% fall at or below 1.00 bit. Table 1 summarizes the empirical scale, and Figure 3 shows the full distribution.

Table 1: Observed entropy distribution in the historical dataset.

Statistic	Entropy (bits)
Minimum	0.326
1st percentile	0.857
5th percentile	1.142
10th percentile	1.283
25th percentile	1.444
Median	1.536
75th percentile	1.570
95th percentile	1.582
99th percentile	1.584
Maximum	1.585

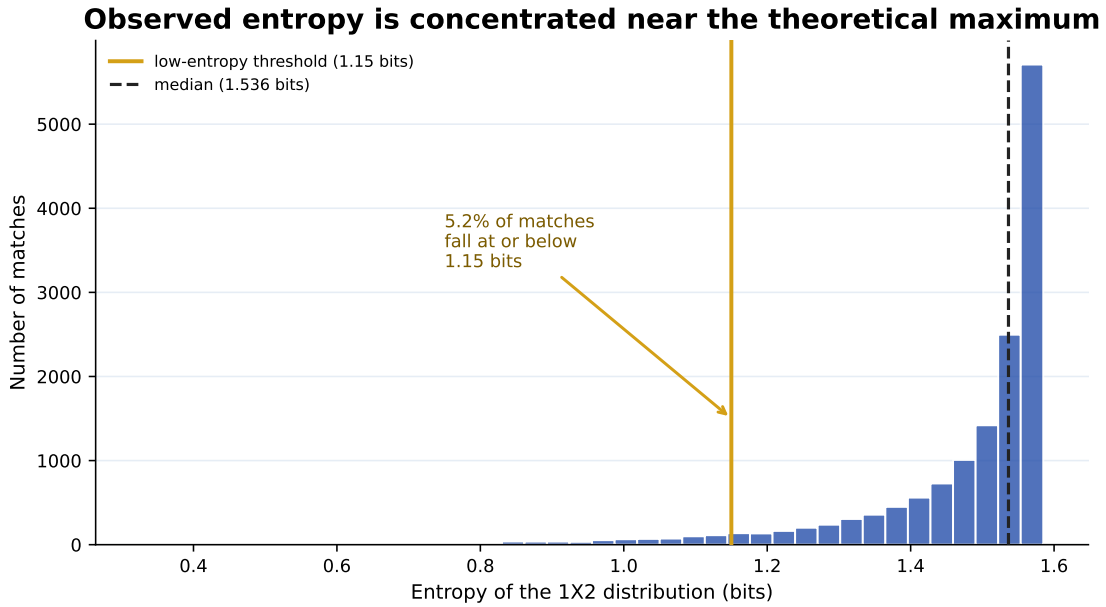

Figure 3: Observed 1X2 entropy distribution. Most matches are close to the high-uncertainty region; the 1.15-bit threshold selects a narrow low entropy subset.

This is a key practical point: $H \leq 1.15$ is not presented as "almost no risk." It is a low entropy region relative to the distribution of calibrated football forecasts.

4 Method

The goal of this analysis is not to train a new forecasting model. Instead, we treat the model output itself as the object of study. Each match already has a calibrated 1X2 probability distribution, and the question is whether an additional distribution level summary can help communicate forecast confidence more responsibly.

This makes the analysis a post-hoc uncertainty filtering problem. Given a probability vector, can we identify forecasts whose full distribution is more concentrated than what $p_{\max}$ alone suggests? The answer is evaluated descriptively on historical outcomes, using aggregate results only.

4.1 Forecast object

Each match is represented by a calibrated 1X2 probability distribution:

$$p = (p_{home}, p_{draw}, p_{away}).$$

The top pick prediction is defined by:

$$\hat{y} = \arg \max_{i \in \{home, draw, away\}} p_i. \tag{4}$$

A prediction is counted as correct if $\hat{y}$ matches the observed 1X2 outcome. This top pick accuracy is used here as an intuitive outcome-level diagnostic, not as the only possible evaluation metric. Proper scoring rules such as log loss, Brier score, or ranked probability score remain better suited for evaluating the full probabilistic forecast. The focus of this note is narrower: how to communicate confidence in the selected top outcome without discarding the rest of the distribution.

4.2 Comparison filters

The first comparison is deliberately naive:

$$p_{\max} \geq 0.60. \tag{5}$$

This threshold represents a forecast that appears confident by headline probability alone. It is not presented as an optimized rule; it is a simple baseline for asking whether $p_{\max}$ is sufficient. In product terms, this is a common intuition: if the most likely outcome exceeds a visible probability threshold, the forecast may look confident. Comparing this rule with an entropy filter allows us to separate two notions that are often conflated: the height of the most likely outcome and the concentration of the full distribution.

The entropy-based filter is:

$$H(p) \leq 1.15 \text{ bits}. \tag{6}$$

This threshold was chosen as a practical compromise: strict enough to isolate a low entropy region, but not so strict that the segment becomes nearly invisible. It should not be interpreted as a universal constant for football. It is a working point observed on this historical dataset, chosen to illustrate the trade-off between selectivity and volume. The more general methodological point is that entropy provides a continuous concentration signal, and any operational threshold should be monitored over time.

We compare:

- the naive high probability baseline $p_{\max} \geq 0.60$;
- the low entropy subset $H(p) \leq 1.15$;
- alternative $p_{\max}$ and entropy thresholds to examine volume-versus-signal trade-offs;
- home favorite and away favorite subsets, because side specific behavior can differ.

4.3 Minimal implementation

The entropy computation is straightforward:

Listing 1: Minimal Python implementation of Shannon entropy in bits.

```python
import numpy as np

def shannon_entropy_bits(p):
    p = np.asarray(p, dtype=float)
    p = p[p > 0]
    return -np.sum(p * np.log2(p))

p = [0.70, 0.25, 0.05]
H = shannon_entropy_bits(p)  # 1.076 bits
is_low_entropy = H <= 1.15
```

5 Results

5.1 Why a naive probability threshold is not enough

A first way to define confidence is to keep only forecasts where the leading probability exceeds a threshold. Table 2 shows the behavior of several $p_{\max}$ thresholds. As expected, increasing

the threshold improves global success rates and reduces volume. However, the recent window behavior is less stable: the $p_{\max} \geq 0.60$ segment records 62/100 on the latest 100 matches, despite a global rate of 79.2%.

Table 2: Naive $p_{\max}$ threshold sweep. Higher $p_{\max}$ improves global rates but reduces volume and does not fully characterize recent uncertainty.

Filter	Volume	Correct	Global success	Latest 100
$p_{\max} \geq 0.50$	5,020	3,504	69.8%	63/100
$p_{\max} \geq 0.55$	3,361	2,502	74.4%	63/100
$p_{\max} \geq 0.60$	2,157	1,709	79.2%	62/100
$p_{\max} \geq 0.65$	1,325	1,114	84.1%	71/100
$p_{\max} \geq 0.70$	800	697	87.1%	83/100
$p_{\max} \geq 0.75$	429	386	90.0%	86/100
$p_{\max} \geq 0.80$	199	178	89.4%	85/100

The entropy filter answers a different question. Instead of asking whether the leading probability is high enough, it asks whether the full distribution is concentrated. Figure 4 compares the naive $p_{\max} \geq 0.60$ filter with $H \leq 1.15$. The entropy subset is much smaller, which is expected: it is a selective filter. But it is also cleaner both globally and in the latest 100 observations.

The recent drop of the naive $p_{\max} \geq 0.60$ segment is not best interpreted as a collapse of the forecasting model. It is better read as a loss of discriminative power for a fixed top probability threshold in a noisy temporal window. The latest 100 forecasts selected by the naive threshold run from 2026-04-26 to 2026-05-11, a period that overlaps with the end of many domestic seasons in the dataset. Such periods can contain asymmetric incentives, rotation, fixture congestion, and favorites that remain plausible on $p_{\max}$ but fragile in distributional terms. This contextual interpretation is not used as a rule in the filter; it is a way to understand why a high-$p_{\max}$ segment can become less reliable while an entropy-based subset remains more concentrated.

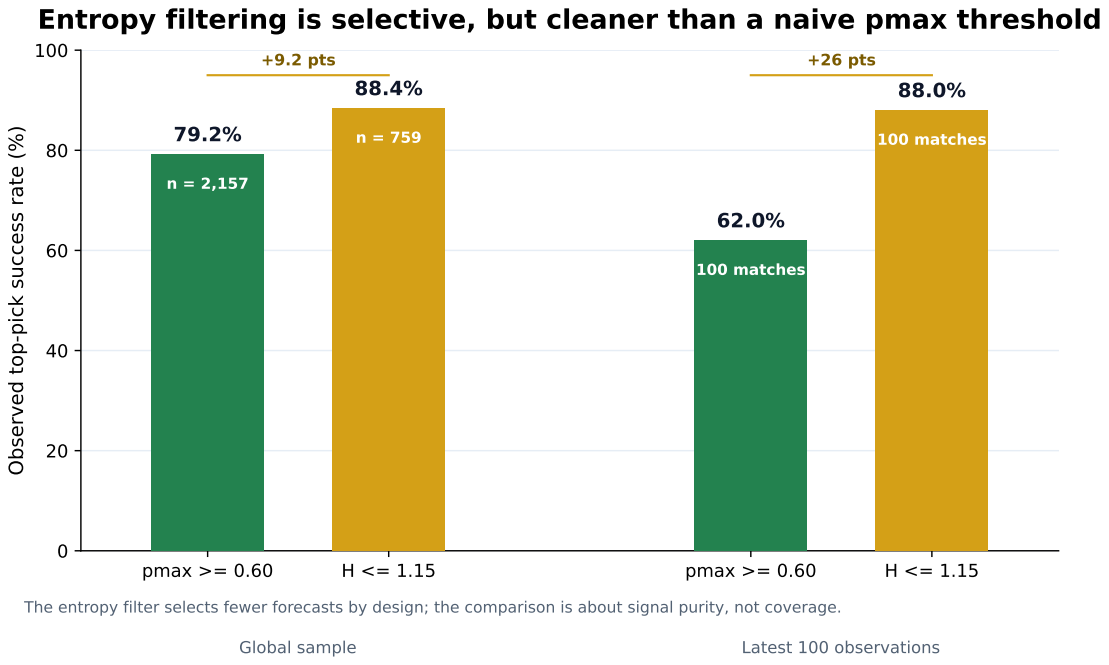

Figure 4: Comparison of a naive top probability filter and the low entropy filter. The entropy subset is selective, but has higher observed success in this historical case study.

The interpretation should remain conservative. The low entropy segment is not simply "better" in a universal sense; it is more selective. The useful result is that entropy identifies a smaller region where the forecast distribution is less diffuse.

The error anatomy of the recent high-$p_{\max}$ segment also supports the draw-trap interpretation. In the latest 100 observations selected by $p_{\max} \geq 0.60$, 38 forecasts are incorrect; 27 of those errors end as draws, and 21 are home favorite forecasts that end as draws. In other words, the recent weakness is largely not caused by the model selecting completely implausible favorites. It is caused by plausible favorites failing to convert, which is precisely the kind of residual uncertainty that $p_{\max}$ alone cannot describe.

Table 3: Error anatomy of the latest 100 forecasts selected by the naive $p_{\max} \geq 0.60$ threshold. The recent weakness is dominated by draw outcomes, especially home favorites ending as draws.

Quantity	Count
Latest 100 forecasts selected by $p_{\max} \geq 0.60$	100
Correct top pick outcomes	62
Incorrect top pick outcomes	38
Incorrect forecasts ending as draws	27
Home-favorite forecasts ending as draws	21

This is why the latest 100 comparison is informative but not volume-neutral. The entropy filter does not recover all high-$p_{\max}$ forecasts. It discards many of them. Its value is that it preserves a smaller subset whose full 1X2 distribution remains concentrated even when the broader probability-threshold segment is diluted by draw-prone favorites.

5.2 Entropy threshold sweep

A threshold analysis helps explain why 1.15 bits is a reasonable working point. Wider thresholds increase volume but dilute the signal, especially in the latest 100 matches. Table 4 gives the detailed sweep, while Figure 5 shows the trade-off graphically.

Table 4: Entropy threshold sweep. Wider thresholds add volume but reduce the recent low entropy signal.

Threshold	Volume	Correct	Global success	Latest 100
$H \leq 1.00$	346	317	91.6%	89/100
$H \leq 1.05$	454	412	90.7%	88/100
$H \leq 1.10$	586	534	91.1%	90/100
$H \leq 1.15$	759	671	88.4%	88/100
$H \leq 1.20$	977	855	87.5%	84/100
$H \leq 1.25$	1,255	1,081	86.1%	76/100
$H \leq 1.30$	1,610	1,357	84.3%	70/100
$H \leq 1.40$	2,816	2,193	77.9%	62/100
$H \leq 1.50$	5,402	3,799	70.3%	64/100

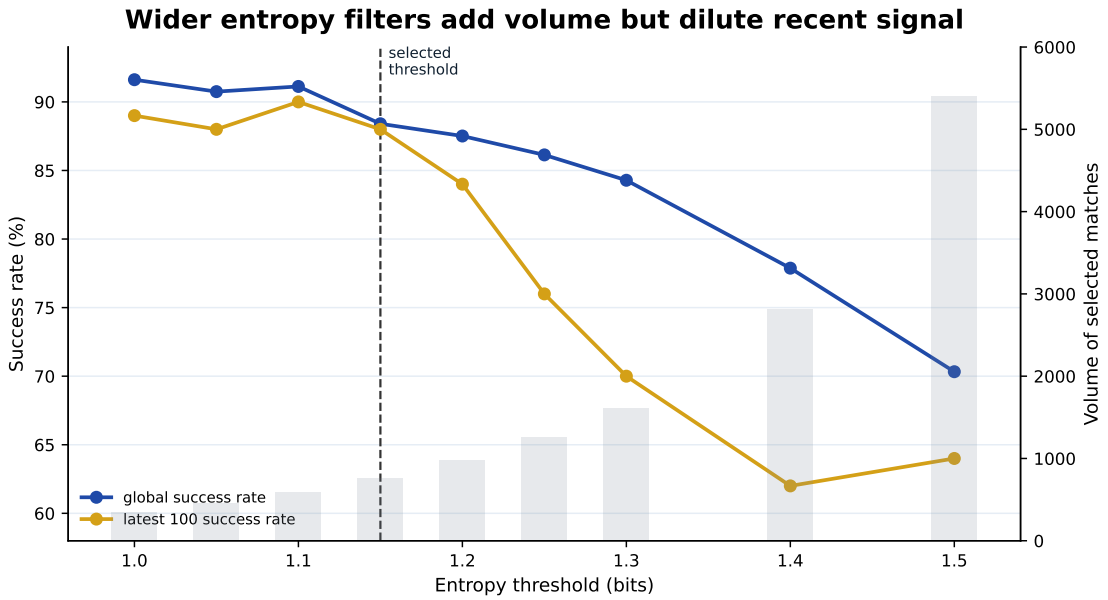

Figure 5: Entropy threshold sweep. The selected threshold balances usable volume with signal purity. Thresholds above 1.20 add volume but dilute recent performance.

This table is also a useful reminder that thresholding is a volume trade-off. A stricter threshold such as $H \leq 1.10$ performs strongly but selects fewer matches. A wider threshold such as $H \leq 1.25$ increases coverage, but the latest 100 signal drops to 76/100. The threshold is therefore not a universal constant; it is an operating point that should be revalidated as the model and seasons evolve.

5.3 Recent windows

Table 5 compares the naive $p_{\max} \geq 0.60$ baseline with the low entropy filter across recent windows. The low entropy subset has lower volume by construction. The point is not to maximize coverage; it is to identify forecasts whose distributions are less diffuse.

Table 5: Recent-window performance. The low entropy filter is more selective but cleaner in the windows shown.

	$p_{\max} \geq 0.60$			$H \leq 1.15$		
Window	Volume	Correct	Rate	Volume	Correct	Rate
7 days	33	24	72.7%	12	10	83.3%
15 days	84	51	60.7%	25	19	76.0%
30 days	168	113	67.3%	47	39	83.0%
90 days	407	278	68.3%	104	92	88.5%
Latest 100	100	62	62.0%	100	88	88.0%

The latest 100 comparison is visually strong, but it must be read with the volume difference in mind. The entropy filter is not a free improvement; it is a selective lens. The recent period is also a useful stress test: it is close to the end of several league seasons, when motivation, squad management, and match context can make apparently strong favorites harder to interpret. Entropy does not model those causes directly, but it can prevent a forecast with residual draw

mass from being communicated as if it were robust.

5.4 Side-specific behavior

Football forecasts often behave differently for home and away favorites. Table 6 and Figure 6 compare home favorite and away favorite subsets. The low entropy filter improves both sides in this historical sample, but away favorites remain weaker than home favorites even after filtering.

Table 6: Home and away favorite subsets. Entropy improves both groups but does not erase side specific difficulty.

	$p_{\max} \geq 0.60$			$H \leq 1.15$		
Segment	Volume	Correct	Rate	Volume	Correct	Rate
Home favorite	1,260	1,031	81.8%	450	407	90.4%
Away favorite	897	678	75.6%	309	264	85.4%

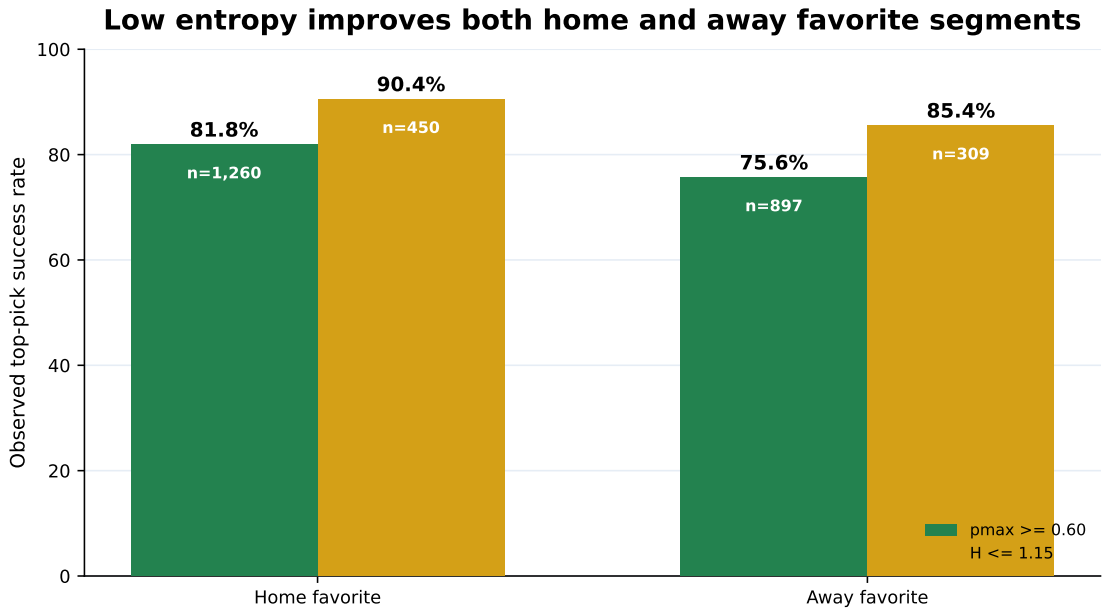

Figure 6: Side-specific behavior. Low entropy improves both home and away favorite segments, but away favorites remain less reliable in this sample.

5.5 Residual error modes

Entropy does not make football deterministic. Incorrect predictions in both the naive high probability baseline and the low entropy subset are still dominated by draws. For $p_{\max} \geq 0.60$, 66.1% of incorrect predictions end as draws. For $H \leq 1.15$, 75.0% of incorrect predictions end as draws.

This result is consistent with the conceptual argument: entropy can identify forecasts with a more concentrated distribution, but it does not remove the structural uncertainty of football. It can help avoid presenting a diffuse forecast as robust; it cannot make the draw disappear.

6 Discussion

6.1 Why $p_{\max}$ and entropy are complementary

The top probability identifies the predicted class. Entropy characterizes how much uncertainty remains across all classes. In a 1X2 football forecast, this distinction is especially relevant because the draw is not merely another low-probability class; it is a frequent and structurally important outcome.

A high $p_{\max}$ and a low entropy often align, but not always. A useful example from the aggregate analysis is the high probability but high-entropy region: forecasts with $p_{\max} \geq 0.60$ and $H > 1.15$ contain 1,399 matches and record 74.3% observed success, while the low entropy subset records 88.4%. In the high-entropy failures, the average draw probability is about 22%, which is large enough to explain why headline confidence can be misleading.

6.2 Low entropy is not low absolute risk

A low entropy football forecast can still carry meaningful absolute risk. The distribution 70/25/5 has entropy of approximately 1.076 bits, which is low for football, but the draw remains 25%. This is why the phrase "confidence signal" is preferable to "safe pick." Entropy measures concentration, not certainty.

6.3 Goal-volatility entropy is a different object

The word "entropy" is sometimes used loosely in sports analytics to describe match volatility, goal openness, expected-goal dispersion, or goal market dynamics. Those ideas can be useful, but they are not the same object as the Shannon entropy of the final 1X2 probability distribution. A high-chaos match may be open in terms of scoring dynamics and still be poor for outcome confidence. Conversely, low 1X2 entropy directly measures concentration in the target outcome space.

6.4 Why the rule is intentionally simple

The low entropy rule is almost boring:

$$H(p) \leq 1.15.$$

That simplicity is a feature. It avoids introducing team-specific, league-specific, or season-specific patches. The filter does not require the model to know that a specific club is dangerous, that a specific league is unusual, or that a specific end of season context is tricky. It simply asks whether the forecast distribution is concentrated. This matters for the recent window interpretation: end of season dynamics may help explain why some favorites failed, but the entropy rule itself remains distributional rather than contextual.

This does not make the threshold universal. It does make the signal easier to explain and easier to monitor.

7 Responsible interpretation and data sourcing

This analysis is retrospective by design. That is appropriate for the question being asked: whether entropy is a discriminating factor in historical forecast outputs. The result should not be read as a guarantee, but as evidence that distributional concentration contains useful information beyond a simple headline probability threshold.

The dataset is proprietary to Foresportia and results from approximately two years of data processing and historical forecast generation. Only aggregate metrics are reported here. No raw match-level dataset is redistributed, and no table of individual forecasts is provided. The article is intended as an applied case study in uncertainty communication, not as a public benchmark dataset.

The main practical limitation is calibration dependence. Entropy is only meaningful if the underlying probabilities are reasonably calibrated. A poorly calibrated model can produce low entropy for the wrong reasons. The threshold should therefore be monitored through rolling evaluation and recalibrated as the forecast system, competitions, and seasons evolve.

Finally, this note does not provide betting advice, financial advice, or individualized prediction recommendations. The figures are aggregate diagnostics of probabilistic forecasts, not match by match decision tools.

8 Conclusion

In multi class sports forecasting, the most likely outcome is not the whole forecast. A high top probability can still hide meaningful uncertainty, especially when the residual mass is concentrated in a structurally important outcome such as the draw.

Entropy provides a simple way to ask a better question:

How concentrated is the full probability distribution?

In this football case study, a naive $p_{\max} \geq 0.60$ filter selected 2,157 forecasts with 79.2% observed historical success. A low entropy filter, $H \leq 1.15$, selected 759 forecasts with 88.4% observed success. The second subset is much smaller; that selectivity is the point. It shows how entropy can help separate forecasts that merely look confident from forecasts whose full distribution is actually concentrated.

The broader lesson is simple: in probabilistic forecasting, confidence should not be inferred only from the top class. It should be read from the shape of the entire distribution.

References

[1] C. E. Shannon. *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. Available at: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

[2] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1950. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[3] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 2007. DOI: 10.1198/016214506000001437

[4] E. Wheatcroft. Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. arXiv:1908.08980, 2019. Available at: https://arxiv.org/abs/1908.08980

[5] Foresportia. Technical Note I - Modeling Football as a Probabilistic Uncertainty Problem. 2026. Available at: https://www.foresportia.com/en/blog/technical-note-1-probabilistic-foot html

[6] Foresportia. Technical Note III - Entropy, Margin and Confidence in Football Predictions. 2026. Available at: https://www.foresportia.com/en/blog/technical-note-3-entropy-confidence html